Term Of References

IF 4090 – Industrial Practice

Organization : National Institute of Technology, Gifu College

Supervisor : Prof. Makoto Yasuda

Student Name : Geraldi Dzakwan

Student ID : 13514065

Informatics Engineering

School of Electrical Engineering and Informatics

Bandung Institute of Technology

2017

TERM OF REFERENCES

Term of References (TOR) contains requirements and works that must be done by the student during the internship program to accomplish the research projects that will be conducted at Yasuda Laboratory in National Institute of Technology, Gifu College (NIT, Gifu College).

1. Objectives

Generally, this internship program aims to develop basic research skill in some specific fields that depends on which laboratory the student will choose and what research topic the student will pick. In addition, this internship program is expected to be a bridge between the two countries, Indonesia and Japan, in order to maintain a good relation, especially between Bandung Institute of Technology (ITB) and NIT, Gifu College. This program is also expected to improve the students' communication skill from both schools especially in English.

Regarding the research subject, the student will be given several research topics from each laboratory. For this term of internship, there are two laboratories which accept ITB students. They are Tajima laboratory and Yasuda laboratory. Then, the students need to choose their preferred laboratory and their preferred topic or they can also propose their own topic (if approved by the laboratory professor). Afterwards, the student will do research project based on the selected topic. The research project will be conducted for 5 to 6 weeks and it will be supervised by the laboratory professor directly. At the end of the internship program, the student shall submit his internship report to the internship supervisor in NIT, Gifu College.

2. Project Scope

The research that will be conducted during the internship program is related to computer science field of study, especially natural language processing and machine learning. The research title is "English Message Anonymization using Trained Named Entity Recognizer". It will be held at Yasuda Laboratory and supervised by Prof. Makoto Yasuda. The student will carry out two projects, which are:

a. Training Named Entity Recognizer

In this project, the student should be able to make a program for predicting named entities from an input word/phrase. The program will use a trained model that will be built using Python Scikit-Learn module (a popular machine learning module in Python). The student will do several experiments using different classifier and/or different data train sample. The student will then decide which model is to be used in the message anonymization API based on the accuracy testing. For the dataset which will be used in training, the student will use Groningen Meaning Bank corpus as so far it is the biggest, most complete, and most suitable English corpus the student has found.

b. Creating Message Anonymization Program

This project is the next step after the first project. This project will use the model which has been created in the first project. The program will probably be in the form of web API, built using either Flask or Django framework. This project will receive an input message (a private message, consisting of one or more sentences, using JSON format) from a client via HTTP POST request. Then, it will do some pre-processing procedures (e.g. sentence tokenizing, POS tagging, NP-chunking and punctuation removal). The result of preprocessing will be delivered to the classifier. The classifier then will return the sentences back to the program with identified named entities. Afterwards, the program will do some post-processing procedures (e.g. correction, co-reference resolution, and sentence restructuring). Last but most important procedure is to apply the anonymization. There are two ways of doing the anonymization, the first one is to replace the named entities with their class (PERSON, LOCATION, ORGANIZATION, etc.) or to replace it with another phrase which belongs to the same class so it becomes more natural. The student will try to explore both options. The restructured sentences which have been anonymized then be sent back to the client via JSON message response.

3. Project Results

The output of this internship program is a functional English message anonymization program as mentioned in the second project of project scope. The message anonymization program will also include the result of the first project as it uses the model produced from the first project. There will also be a final report in the form of printed document and a final presentation file to be delivered at the end of the internship program.

4. Project Constraints

The followings are some constraints that are applied throughout the first project about training named entity recognizer:

1. The programming language used is Python version 2.7.10.
2. The English corpus dataset used is Groningen Meaning Bank (GMB) corpus version 2.2.0.
3. The main machine learning tools used for training the model is Python Scikit-Learn version 0.17.
4. The classes of named identity that can be identified are limited to: person name, location/geographical entity, organization, geopolitical entity, and time indicator.
5. Accuracy testing will be done by split test, which is separating train dataset and test dataset from the corpus.

The followings are some constraints that are applied throughout the second project about creating message anonymization program:

1. The programming language used is Python version 2.7.10.
2. The main tools for doing natural language processing techniques are Python NLTK (Natural Language Toolkit) version 3.2.4.
3. The framework used for creating web API is either Flask or Django.
4. The program will only receive private message (a message that has to be anonymized) as the input.
5. Testing will be done by data from GMB corpus and other data that are collected manually by the student.

5. Assumptions

The followings are some assumptions regarding status or conditions which are agreed for conducting the first project:

1. The student has basic knowledge of computer science and programming.
2. The student is proficient in using Python programming language.
3. The student has basic knowledge of machine learning, knowing how to do training and testing of dataset.
4. The student is eager to learn about Python Scikit-Learn tools to do the machine learning processes.
5. The project can be done within 2 weeks.

The followings are some assumptions regarding status or conditions which are agreed for conducting the second project:

1. The student has basic knowledge of computer science and programming.
2. The student is proficient in using Python programming language.
3. The student has basic knowledge of natural language processing, knowing how to extract information from raw text (tokenizing sentences, giving POS tag, chunking noun phrase, etc.).
4. The student is eager to learn about Python NLTK tools to do the natural language processing procedures and also to learn the web API framework.
5. The project can be done within 3 weeks.

6. Work Methodology

In order to successfully finish the research project, there are some strategies, including the method, process and/or tools, which are summarized in a work methodology of the projects below:

a. Training Named Entity Recognizer (1st Project)
    1. Set up the work environment for the first project, including tools, libraries, and other dependencies that should be installed.
    2. Learn about the structure of the corpus.
    3. Learn how to create and train classifier in Python Scikit-Learn.
    4. Program development.
    5. Model training and testing.
    6. Analysing the best classifier and model to use.
    7. Make report.

b. Creating Message Anonymization Program (2nd Project)
   1. Set up the work environment for the second project, including tools, libraries, and other dependencies that should be installed.
   2. Learn how to do common NLP tasks (tokenizing sentences, POS tagging, NP-chunking, etc.) using Python NLTK.
   3. Learn how to create web framework API using Django or Flask.
   4. Program development.
   5. Collecting data test.
   6. Program testing.
   7. Analysing the accuracy and program performance.
   8. Improve accuracy and performance.
   9. Make report.

c. Documentation
   1. Make final report and presentation.

7. Development Environment

The following information is the environment details in developing the program:

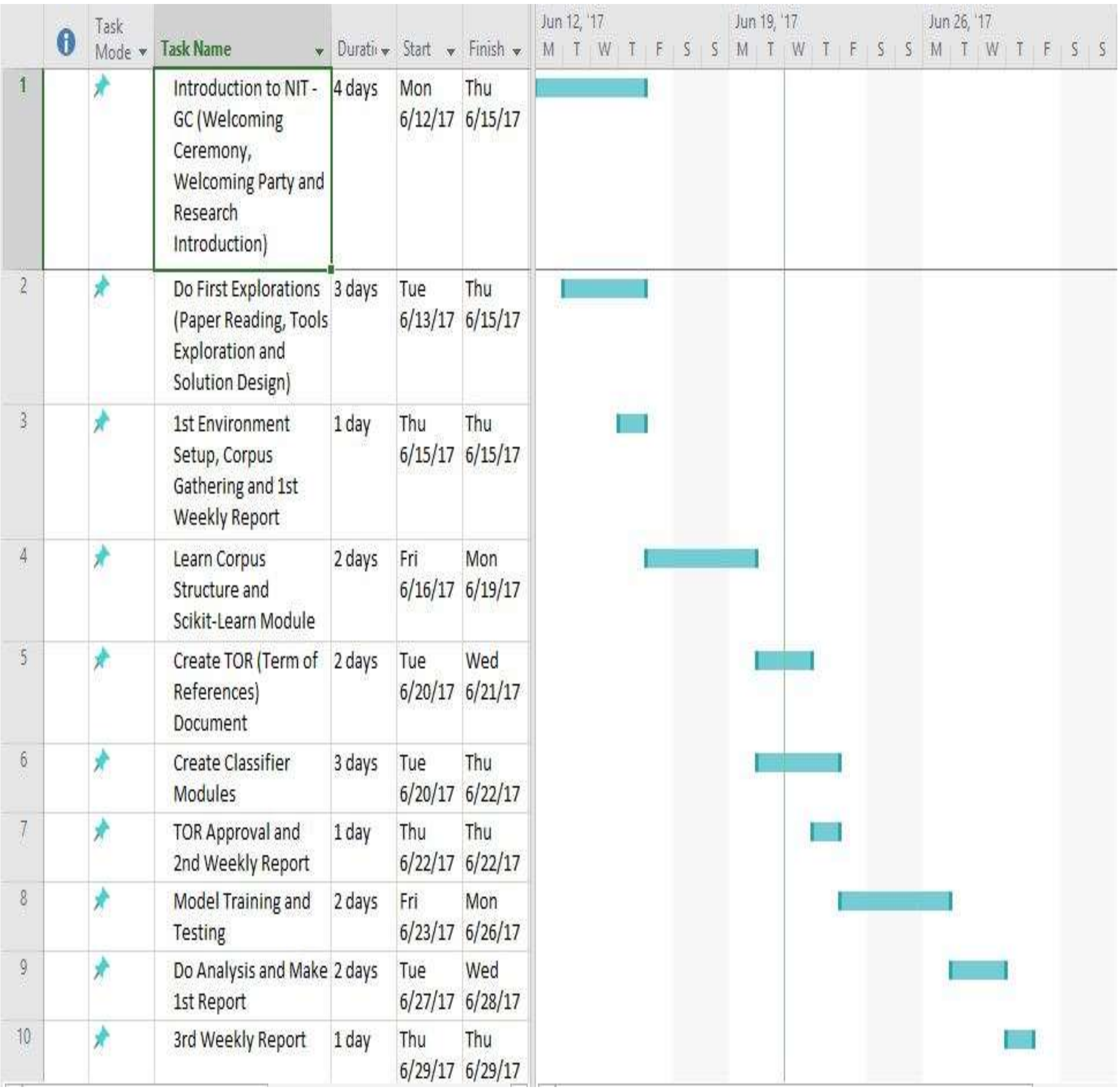
a. Hardware Specifications
   1. Operating System : Ubuntu LTS 16.04
   2. Processor : Intel® Core™ i5-4200U @ 1.6GHz
   3. RAM : 12.00 GB DDR3 1600MHZ
   4. System Type : 64-bit operating system, x64-based processor

b. Software Specifications
   1. Training Named Entity Recognizer (1st Project)
      a. Programming Language : Python version 2.7.10
      b. Requirements : Python Scikit-Learn module
      c. IDE : Atom version 1.18.0 for Ubuntu

   2. Creating Message Anonymization Program (2nd Project)
      a. Programming Language : Python version 2.7.10
      b. Requirements : Python NLTK module
      c. IDE : Atom version 1.18.0 for Ubuntu
      d. Web Framework : Django or Flask

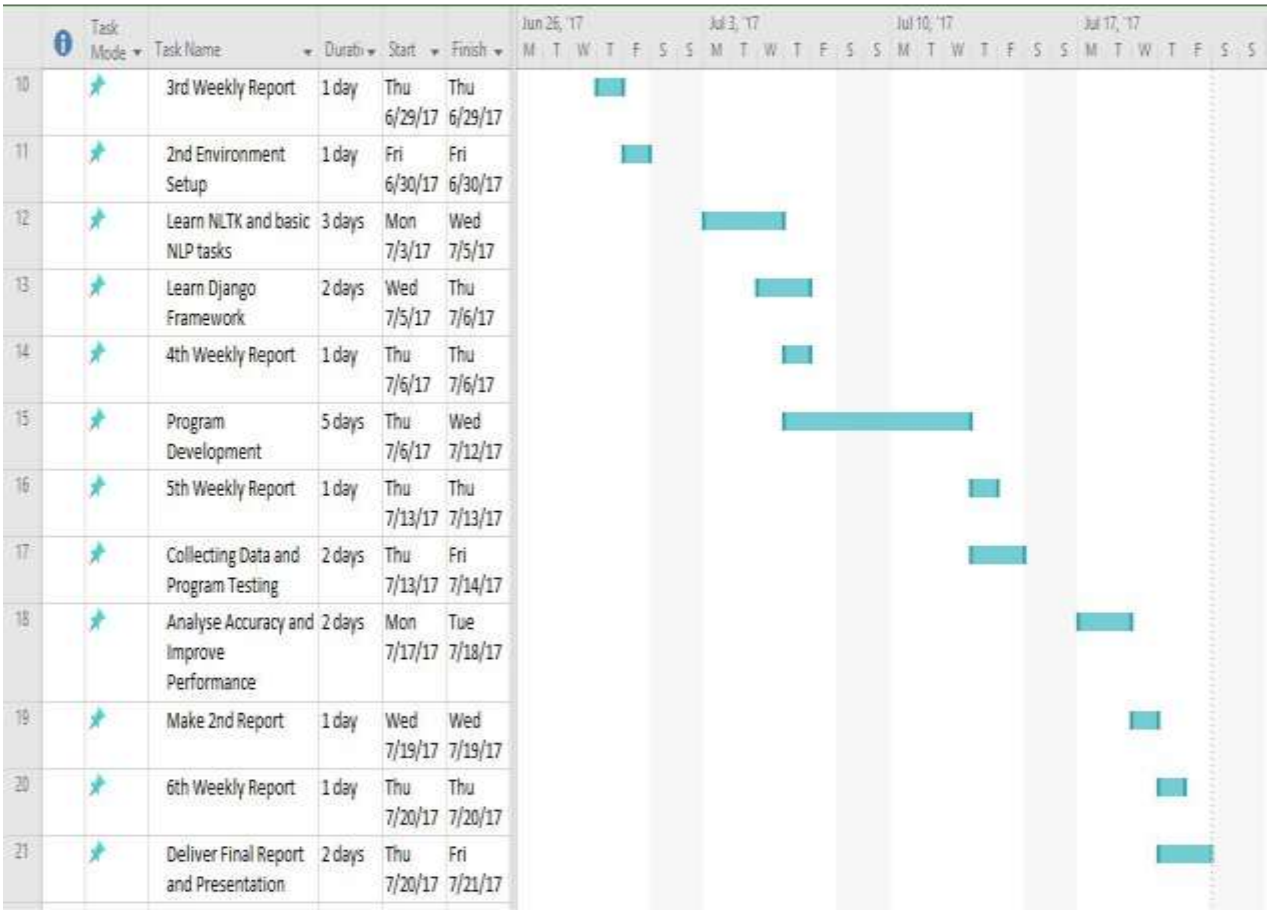
8. Schedule

Generally, the student must come to the laboratory every work day (Monday-Friday), from 9.00 a.m. until 16.10 p.m. to work on the research project. Despite that, other activities might take place during research time, such as attending welcoming ceremony and welcoming party, weekly talk café (interacting with NIT-GC students) and giving presentation in some classes. The detailed schedule can be seen below. There are two pictures, the first one shows Gantt chart for the first project and the second one shows Gantt chart for the second project.



	ℹ	Task Mode	Task Name	Durati	Start	Finish
1		📌	Introduction to NIT - GC (Welcoming Ceremony, Welcoming Party and Research Introduction)	4 days	Mon 6/12/17	Thu 6/15/17
2		📌	Do First Explorations (Paper Reading, Tools Exploration and Solution Design)	3 days	Tue 6/13/17	Thu 6/15/17
3		📌	1st Environment Setup, Corpus Gathering and 1st Weekly Report	1 day	Thu 6/15/17	Thu 6/15/17
4		📌	Learn Corpus Structure and Scikit-Learn Module	2 days	Fri 6/16/17	Mon 6/19/17
5		📌	Create TOR (Term of References) Document	2 days	Tue 6/20/17	Wed 6/21/17
6		📌	Create Classifier Modules	3 days	Tue 6/20/17	Thu 6/22/17
7		📌	TOR Approval and 2nd Weekly Report	1 day	Thu 6/22/17	Thu 6/22/17
8		📌	Model Training and Testing	2 days	Fri 6/23/17	Mon 6/26/17
9		📌	Do Analysis and Make 1st Report	2 days	Tue 6/27/17	Wed 6/28/17
10		📌	3rd Weekly Report	1 day	Thu 6/29/17	Thu 6/29/17

Figure 1: Gantt Chart for 1st Project



	Task Mode	Task Name	Durati	Start	Finish
10		3rd Weekly Report	1 day	Thu 6/29/17	Thu 6/29/17
11		2nd Environment Setup	1 day	Fri 6/30/17	Fri 6/30/17
12		Learn NLTK and basic NLP tasks	3 days	Mon 7/3/17	Wed 7/5/17
13		Learn Django Framework	2 days	Wed 7/5/17	Thu 7/6/17
14		4th Weekly Report	1 day	Thu 7/6/17	Thu 7/6/17
15		Program Development	5 days	Thu 7/6/17	Wed 7/12/17
16		5th Weekly Report	1 day	Thu 7/13/17	Thu 7/13/17
17		Collecting Data and Program Testing	2 days	Thu 7/13/17	Fri 7/14/17
18		Analyse Accuracy and Improve Performance	2 days	Mon 7/17/17	Tue 7/18/17
19		Make 2nd Report	1 day	Wed 7/19/17	Wed 7/19/17
20		6th Weekly Report	1 day	Thu 7/20/17	Thu 7/20/17
21		Deliver Final Report and Presentation	2 days	Thu 7/20/17	Fri 7/21/17

Figure 2: Gantt Chart for 2nd Project

9. Staffing

No staffing needed for this internship because the student will carry out the research by himself which means no other student/party will be involved in the research project.

10. Control and Monitoring Mechanism

While doing the research projects, the student will be monitored by Prof. Makoto Yasuda as the laboratory professor. The professor also acts as the supervisor for the research projects. The followings are further details of control and monitoring mechanism for the internship program:

1. Lab activities begins from 09.00 a.m. until 16.10 p.m. for every work day (Monday-Friday). The student will be monitored by the internship supervisor at NIT, Gifu College (Prof. Makoto Yasuda) during working in his projects.

2. Every Thursday afternoon, there will be weekly report from the student regarding the progress he made and the internship supervisor (Prof. Makoto Yasuda) will evaluate it accordingly.

3. Every Sunday afternoon, the student will send an email to his professor at ITB to summarize the work that has been done within that week (in the form of log activity document).

4. There are several milestones for the project works, listed below:
   a. June 13, 2017 : The student has settled his research topic and project scope for his internship
   b. June 15, 2017 : The student has completed doing his explorations (reading papers, exploring the tools from books and designing solution)
   c. June 22, 2017 : The TOR (term of references) document has been approved by the internship supervisor (Prof. Makoto Yasuda)
   d. June 28, 2017 : The student has successfully finished the 1st project
   e. July 19, 2017 : The student has successfully finished the 2nd project
   f. July 21, 2017 : All results (final program, report and presentation) have been delivered.

5. At the end of the internship program, the student will give a final report and final presentation about the research projects.

11. Risk Management

In case there are works that aren’t done within the internship duration at Japan, the student is willing to continue the work in his home country, Indonesia and finish it within at most one month after departure from Japan.

12. Other Information

The student will receive JASSO scholarship for covering his daily expenses in Japan. The amount of scholarship granted is 80000 JPY, given once every three weeks. So, the total amount would be 160000 JPY as the internship program takes 6 weeks to complete.

Approved by,

Prof. Makoto Yasuda
Internship Supervisor

Geraldi Dzakwan
Internship Student